

Question 1:

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The optimal value for alpha for ridge is 10.

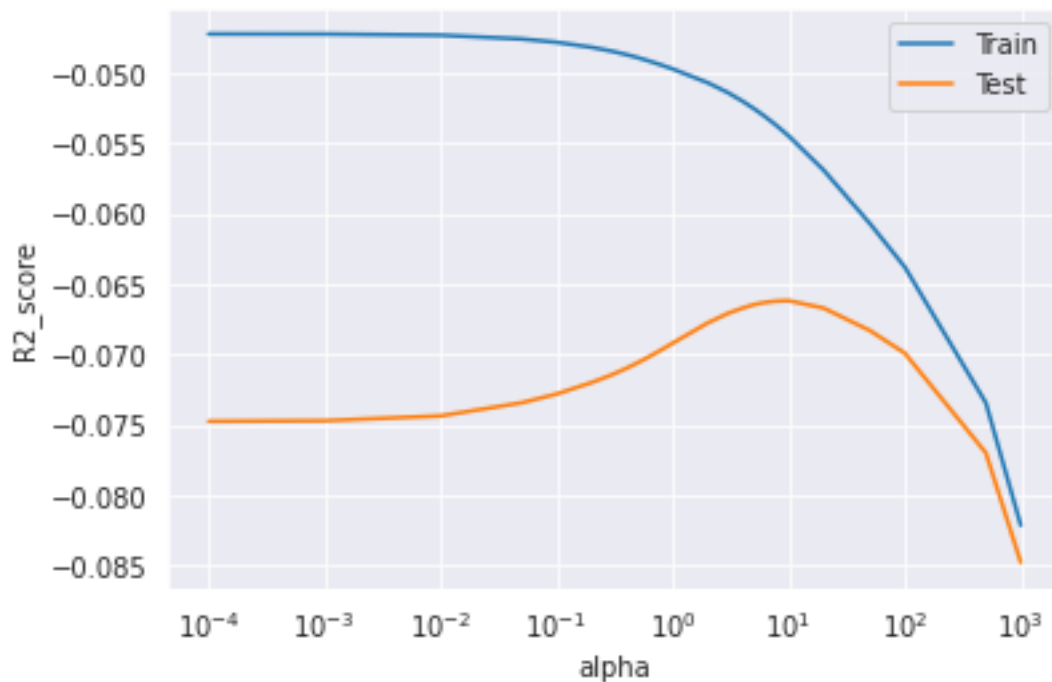
The optimal value of alpha for lasso is 0.001.

In Ridge regression, after doubling the value of the alpha (alpha = 100)

	Metric	Ridge - Original	Ridge - Doubled
0	R2 Score (Train)	0.94	0.92
1	RSS (Train)	4.65	6.23
2	MSE (Train)	0.01	0.01
3	RMSE (Train)	0.07	0.09
4	R2 Score (Test)	0.88	0.86
5	RSS (Test)	2.41	2.82
6	MSE (Test)	0.01	0.01
7	RMSE (Test)	0.11	0.11

- Train R2 score has dropped for the new model.
- Train RMSE has increased.
- Test R2 score has dropped for the new model.
- Test RMSE remains same.

This can be explained by the chart below



As the value of alpha increases, we see a decrease in train error and an initial increase followed by decrease in test error.

From the graph also it is visible that the optimum value for alpha is 10.

Important predictor variables in ridge after the change:

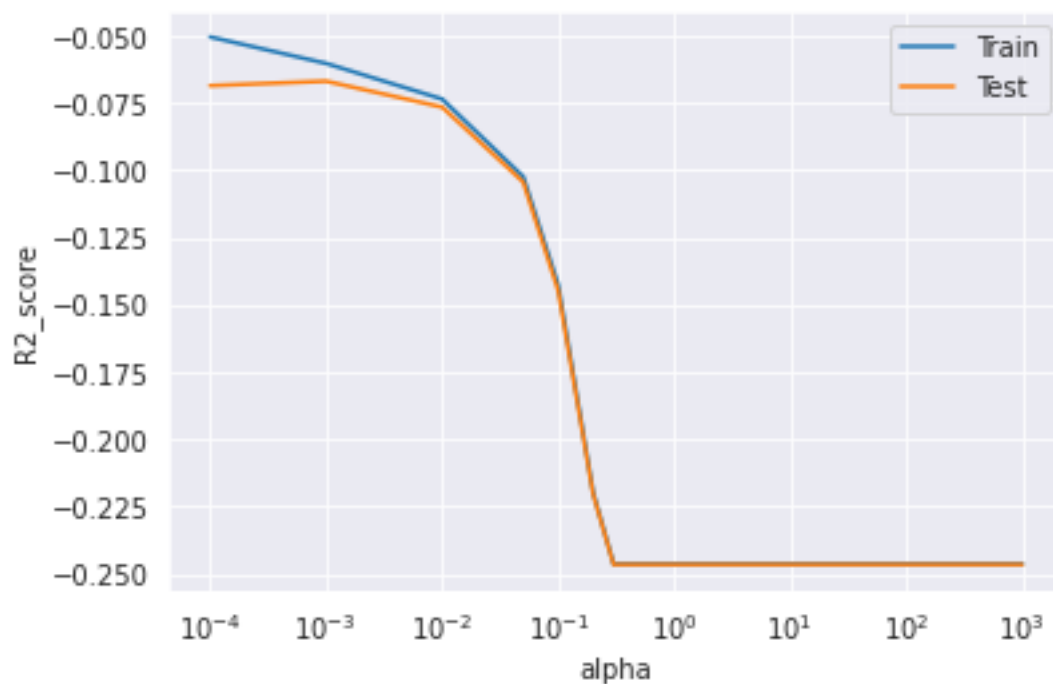
- OverallQual
- GrLivArea
- YearBuilt
- TotalBsmtSF
- OverallCond
- 2ndFlrSF
- Condition1_Norm
- BsmtFinSF1
- 1stFlrSF
- Neighborhood_Crawford

In Lasso regression, after doubling the value of the alpha ($\alpha = 0.000001$)

	Metric	Lasso - Original	Lasso - Doubled
0	R2 Score (Train)	0.93	0.95
1	RSS (Train)	5.67	3.77
2	MSE (Train)	0.01	0.00
3	RMSE (Train)	0.08	0.07
4	R2 Score (Test)	0.86	0.87
5	RSS (Test)	2.65	2.52
6	MSE (Test)	0.01	0.01
7	RMSE (Test)	0.11	0.11

- Train R2 score has increased for the new model.
- Train RMSE has dropped.
- Test R2 score has increased for the new model.
- Test RMSE remains same.

This can be explained the chart below,



As the value of alpha increases, we see a decrease in both train and test error.

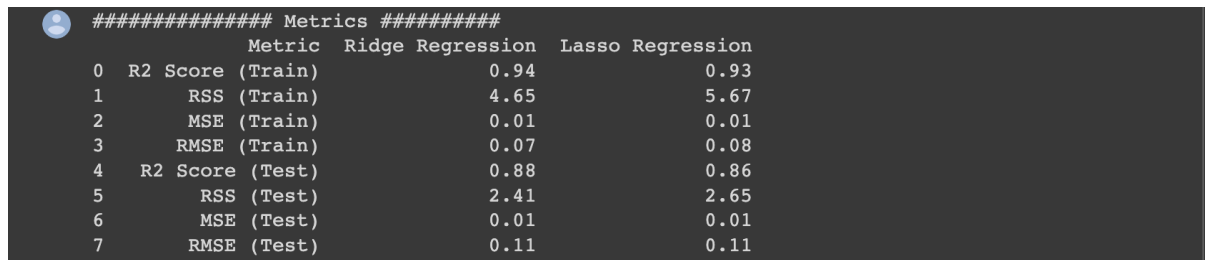
Important predictor variables in Lasso after the change:

- Heating_GasW
- GarageType_None
- Heating_GasA
- Exterior2nd_Stucco
- MSZoning_FV
- MSZoning_RL
- Heating_Grav
- Neighborhood_Crawfor
- MSZoning_RH
- Condition1_PosA

Question 2:

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Comparing both the models.



```
##### Metrics #####
Metric Ridge Regression Lasso Regression
0 R2 Score (Train) 0.94 0.93
1 RSS (Train) 4.65 5.67
2 MSE (Train) 0.01 0.01
3 RMSE (Train) 0.07 0.08
4 R2 Score (Test) 0.88 0.86
5 RSS (Test) 2.41 2.65
6 MSE (Test) 0.01 0.01
7 RMSE (Test) 0.11 0.11
```

Both models have similar train and test scores.

Our model selection will depend on the use case,

- If we need feature selection, we will use Lasso
- If we are ok with including all the parameters in the mode, we will use Ridge.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

- Drop the five columns in the train data.
- Do grid search again to get the optimum value.
- Build the model again with new alpha value.
- Check the coefficients to find the new important variables.

Five important predictor variables after dropping the initial five predictor variables.

- 2ndFlrSF
- 1stFlrSF
- MSZoning_FV
- Functional_Typ
- Foundation_PConc

(P.s: The code changes are available in the notebook)

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

A model is robust when it can produce a consistent predictor variable even when one or more of the input variables change drastically in the unseen test data.

A model is generalizable when it is able to adapt to new, unseen data which is drawn from the same distribution as the one used to create the model.

We make sure a model is generalisable by creating test/train split and by using the test data on the model built with training set and check the model metrics such as R2 score to look for overfitting, under fitting etc.

Regularisation techniques such as Lasso and Ridge also helps in building a generalisable model.

In terms of accuracy,

- A complex, overfitting model will have high test R2 score but will fail in the test data. In such models, the bias will be low but the variance will be high, leading to higher total errors and less accuracy.
- A simple, underfitting model will have low model complexity, however, in such models bias will be high but variance will be low, leading to higher total errors and less accuracy.

So we need to build a model that has optimum bias and variance, we need to strike a balance between accuracy and complexity to build a robust, generalisable model.